B-8 Genome Informatics of Bacteria : Statistical Analysis(基礎物理学研究所研究会「複雑系6」報告,研究会報告)

AUTHOR(S):

福島, 敦史

CITATION:

福島, 敦史. B-8 Genome Informatics of Bacteria : Statistical Analysis(基礎物理学研究所研究会「複雑系6」報告,研究会報告). 物性研究 2000, 74(1): 82-86

ISSUE DATE:

2000-04-20

URL:

http://hdl.handle.net/2433/96797

Genome Informatics of Bacteria

Statistical Analysis

日大理工　福島 敦史 *

1　はじめに

近年、ヒトゲノム計画に代表されるプロジェクトによって、様々な種のゲノム塩基配列のデータが蓄積されている。ゲノムは生物が生きるのに必要な遺伝情報の一揃いのセットのことをいう。ヒトゲノムでは遺伝子のコード領域がゲノム全体の数%であり、遺伝子ではない領域がゲノム配列の持つ本質的な情報の意味や性質を担っているといえる。このようなゲノムのグローバルな性質を調べる研究は、医学的な研究に代表される遺伝子ごとの性質を調べるという研究と同様に重要であると考えられる。

本研究では、バクテリアのcompleteなゲノム塩基配列を、統計物理学的な手法により解析した。ヒトゲノムに対しバクテリアゲノムはそのほとんどが遺伝子、つまりコード領域である。ヒトDNA塩基配列について統計物理学的解析を行った*Stanley*らの主張[1,2]との比較検討を行い、そこからゲノムに関してより詳細で深い理解を得ようとするのが本研究の目的である。その結果、バクテリアゲノムはイントロン(非コード領域)がなく、全体がコード領域の集まりであるにもかかわらず、長距離相関が存在していた。それはタンパク質をコードする配列の長さよりもはるかに長い配列、そのゲノム全長にわたることが明らかとなった。

2　解析結果

解析は10種のバクテリアゲノムデータに関して行なっている。ここでは、塩基配列データが580073塩基のマイコプラズマ菌(Mycoplasma genitalium)に関して説明する。

2.1　G+C含量解析の結果

ゲノムの塩基組成は「G+C%」(配列中に含まれる塩基GとCの割合)として表す。これは2本鎖DNAの場合、化学的にA%=T%、G%=C%が成立しているので、「G+C%」という量を知ることで、2本鎖DNA分子の塩基組成を一義的に記述することができるからである[3]。以上のことから「G+C%」なる量に着目して解析を行なった。

図1はマイコプラズマ菌(Mycoplasma genitalium)の塩基配列データ(570083塩基)をウィンドウサイズL=1000として、第1番目の塩基から1000+1番目の配列中のG+C%を

* E-mail: fuku@phys.cst.nihon-u.ac.jp

求め、第2番目から1000+2番目、…、i番目から1000+i番目という風に順次G+C%を求めて、ゲノム上の位置iに対する$(G+C)_i$%の変動を図示したものである。ゲノム全体の平均GC%は31.7%であった。

図2はウィンドウサイズLに対するG+Cの個数の確率分布で、データから得られた分布に最も近い(分散を一致させている)ガウス分布と比較している。縦軸はlogスケールである。L=100,500,1000のときの分布を示す。

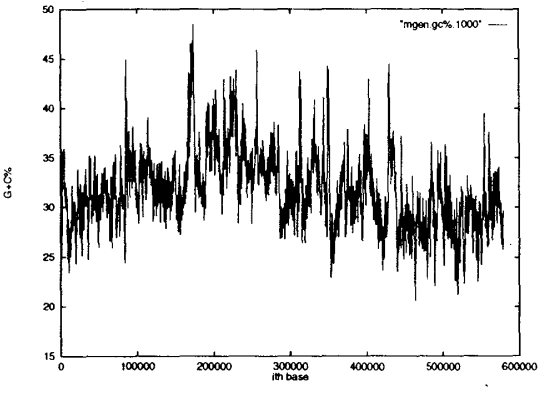

図 1: ゲノム上の位置 i に対する G+C%の変動

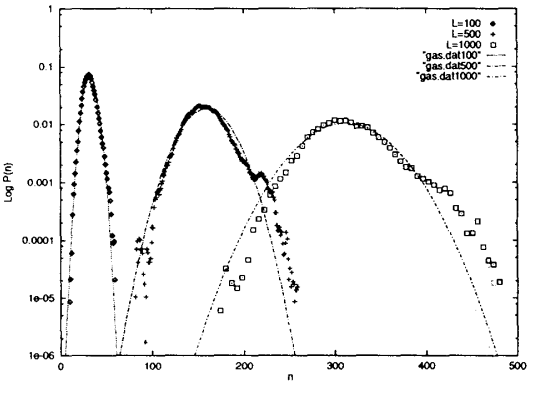

図 2: 長さLの配列中のG+Cの個数nに対する分布 (L=100,500,1000)

図3はL=1000,3000,5000のときの、図4はL=5000,7000,10000のときの分布を示す。Lが大きくなるにつれて裾野がガウス分布からずれてゆくのがわかる。

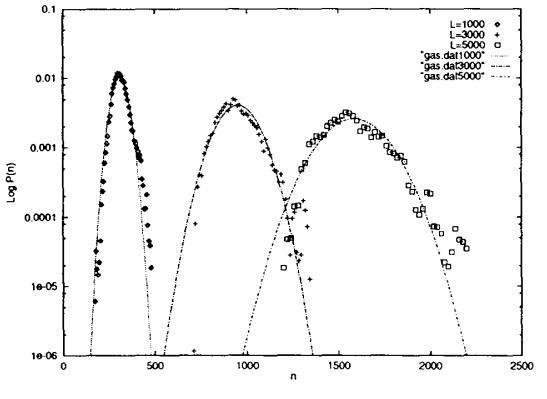

図 3: L=1000,3000,5000 の場合の G+C 個数 n の分布

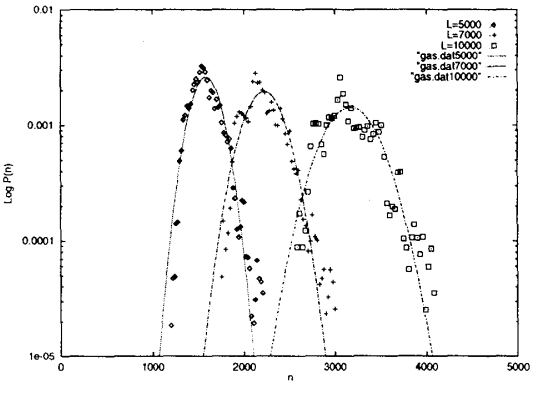

図 4: L=5000,7000,10000 の場合の G+C 個数 n の分布

図5は横軸がL(サンプリングする配列の長さ)で、それに対するG+Cの個数の分布の分散である。これを見ると、L=100を越えたあたりからL=100000まで$L^{1.7}$に乗っている。これは細菌の遺伝子の1つ(約1000塩基から成る)分よりはるかに長く相関が存在するということである。

解析の結果、ほとんどすべてのバクテリアゲノムで、G+Cの個数の分散のウィンドウサイズ依存は$\sim L^{\alpha}$で指数αは1.0よりも大きい値を示すことがわかった。

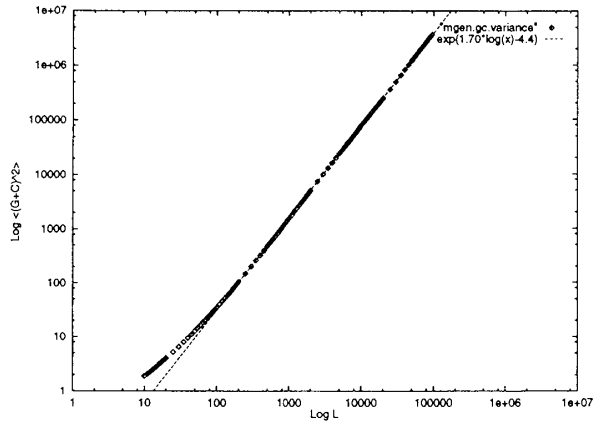

図 5: 配列の長さ L に対する G+C の個数の分散

2.2 仮想ブラウン粒子を用いた解析の結果

1 次元 x 軸上を速度 v で運動する粒子を考え、塩基 G、C が現われたときに粒子の速度を反転させる。そして、$\Delta x = x(t+\Delta t) - x(t)$ として Δt 秒ごとの変位 Δx の分布を求めた。ここで、Δt とは塩基の長さ (ウィンドウサイズ) に相当する。GC の出現事象がポアソン過程に従う場合、仮想ブラウン運動の変位の分布は漸近的にガウス分布に従うことが知られているので、それからのずれを調べる。

図 6 はマイコプラズマ菌 (Mycoplasma genitalium) の塩基配列データ (570083 塩基) による仮想ブラウン粒子の軌道 ($t - x$ 図) である。

図 7 は Δt を変えていったときの変位 Δx に対する分布の分散の変化である。このとき分散は $\Delta t^{1.0}$ で変化していき、ブラウン運動的であり、G+C 含量の解析で現れた相関は新しい座標 Δx では消えている。

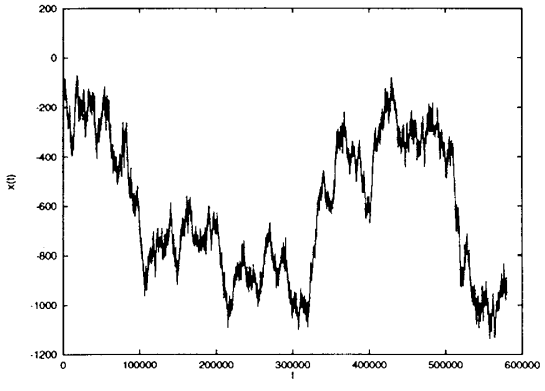

図 6: 塩基配列データによる仮想ブラウン粒子の軌道

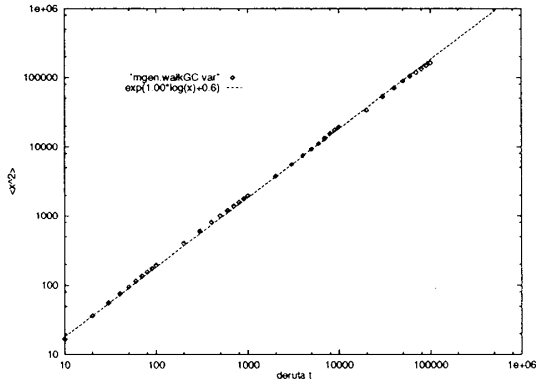

図 7: Δ t に対する変位 Δ x の分散

図8～図11はそれぞれ、$\Delta t = 7000, 8000, 9000, 10000$に対する変位$\Delta x$の分布である。これを見ると$\Delta t = 8000$まではガウス分布であるが、$\Delta t = 9000$を境にガウス分布から大きくずれることがわかる。$\Delta t = 9000$以上は両側指数分布によって*Fitting*している。

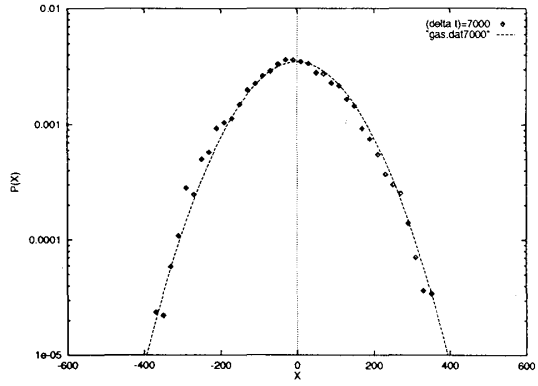

図 8: Δ t=7000 に対する変位Δ x の分布

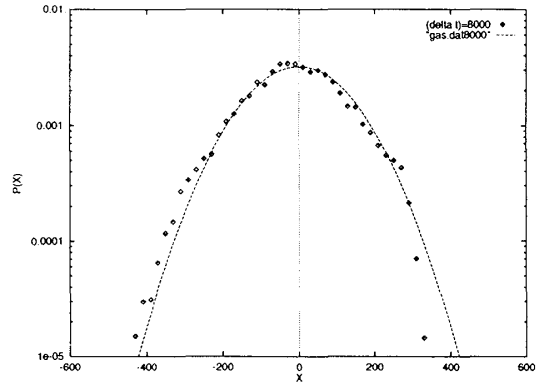

図 9: Δ t=8000 に対する変位Δ x の分布

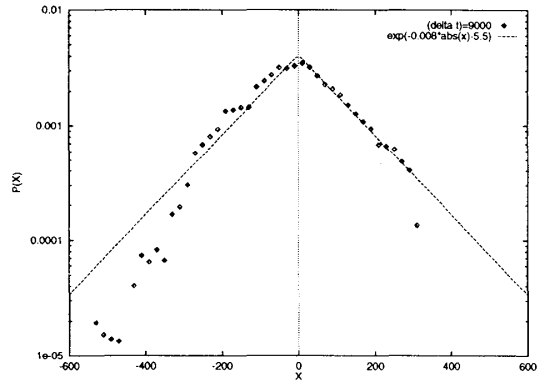

図 10: Δ t=9000 に対する変位Δ x の分布

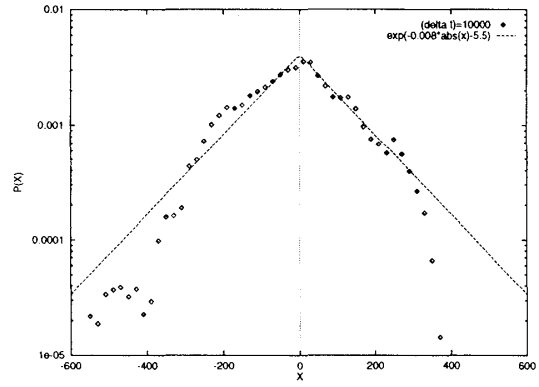

図 11: Δ t=10000 に対する変位Δ x の分布

3 考察と将来展望

現在までにDNA塩基配列の統計物理学的な解析は*Stanley*ら多数の人々によって行なわれている。*Stanley*らのヒトのDNA塩基配列の解析からの主張は非コード領域(イントロン)とそれを含む遺伝子中のDNA塩基配列には長距離相関があり、コード領域(エクソン)において長距離相関は見出されない、ということである。

*Stanley*らの解析したDNA配列データは主に遺伝子単位、あるいは遺伝子内についてのものであった。現在では使用できるデータの量は当時に比べてはるかに多い。現在までにゲノムの全配列が決定された生物種(バクテリア)のデータを解析した本研究の結果によれば、全体がコード領域の集まりであるバクテリアゲノムにおいて、ヒトゲノムのコード

領域には存在しないとされている長距離相関が存在していた。そしてその長距離相関は、ゲノム全長にわたるようなRangeで存在することが示された。さらに、現在まで解析の終わっているバクテリアのすべてにおいて長距離相関が見られ、ゲノムはフラクタル性を持っていると考えられる。現在のところ、ゲノムの持つこのようなフラクタル性(長距離相関)の起源は、理解できていない。

近年、ヒトゲノムに代表される温血脊椎動物のゲノム中に、GC含量のモザイク的な区分構造の存在が示されている[4][5]。そのミクロなGC含量の変動とマクロな構造である、染色体バンドとを対応付けする試みも報告されている。ランダムプロセスである突然変異が、ゲノムの長いRangeにわたる、ある種の秩序を生み出し、かつその構造を進化的に保持していることは非常に興味深いことである。

フラクタル性の起源を知る上で、GC含量そのものが重要であるのか、あるいは特定の塩基の連なりに本質的な意味があるのかは自明ではない。したがって、GC含量そのものの生物学的な意味付けをより厳密に行う必要があり、ゲノムDNA自体の構造や核内における折り畳み構造との詳細な対応付けをすることが重要であるように思う。

また一口に長距離相関といってもその性質は様々であろう。現在までにさまざまなシステムにおける長距離相関をクラス分けする試みも進められている[6]。生物のゲノムは、どのような性質の長距離相関を持つのであろうか。

参考文献

[1] H.E.Stanley, S.V.Buldyrev, A.L.Goldberger, S.Havlin, R.N.Mantegna, C.-K.Peng, M.Simons and M.H.R.Stanley,"*Long-Range Correlations andGeneralized Lévy Walks in DNA Sequences*", Springer,(1994),331-347

[2] S.V.Buldyrev, N.V.Dokholyan, A.L.Goldberger, S.Havlin, C.-K.Peng, H.E. Stanley ,G.M.Viswanathan, "Analysis of DNA sequences using methods of statistical physics", Physica A,(249),1-4,(1998),430-438

[3] 宮田 隆・五条堀 孝 編 ゲノム情報を読む 共立出版(1997)

[4] S.Aota and T.Ikemura, "Diversity in G+C content at the third position of codons in vertebrate genes and its cause", Nucleic Acids Res.,14,(1986),6345-6355

[5] T.Ikemura, K.Wada, and S. Aota, "Giant G+C% Mosaic Structures of the Human Genome Found by Arrangement of GenBank Human DNA Sequences According to Genetic Positions", GENOMICS, 8,(1990),207-216

[6] 五味 壮平 「複雑系5」研究会報告, 物性研究,68-5,(1997),640-648